

DOCUMENT RESUME

ED 174 643

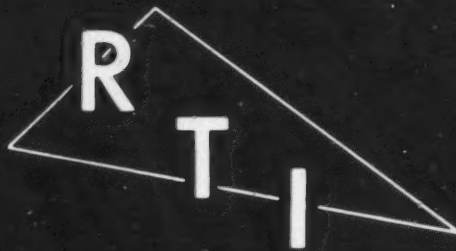
TM 009 265

AUTHOR Wisenbaker, Joseph M.; Schmidt, William H.
TITLE The Structural Analysis of Hierarchical Data.
SPONS AGENCY Michigan State Univ., East Lansing. Inst. for
Research on Teaching.; National Inst. of Education
(DHEW), Washington, D.C.
PUB DATE Apr 79
CONTRACT 400-76-0073
NOTE 19p.; Paper presented at the Annual Meeting of the
American Educational Research Association (63rd, San
Francisco, California, April 8-12, 1979)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Covariance; *Data Analysis; *Multiple
Regression Analysis; Research Design; Senior High
Schools; Statistical Analysis
IDENTIFIERS *Hierarchical Data; LISREL Computer Program; National
Longitudinal Study High School Class 1972

ABSTRACT

The problems inherent in analyzing data in which subjects are "nested" within hierarchical units (such as classrooms or schools), and thus are not independent of one another, are addressed through the use of Joreskog's LISREL model for analyzing covariance matrices. A solution to the problem is proposed and illustrated using data from the National Longitudinal Study of the High School Class of 1972. (JKS)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *



U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

RESEARCH TRIANGLE INSTITUTE

The Structural Analysis of Hierarchical Data

Joseph M. Wisenbaker
Research Triangle Institute
and

William H. Schmidt
Michigan State University

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Joseph M. Wisenbaker

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

A paper presented at the American Educational Research Association
Annual Meeting

April 8-12, San Francisco, California

Session # 29.03

The work reported herein is sponsored by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract Number 400-76-0073)

RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

ED174643

TM009 265

The type of data considered in this paper is one which poses special problems for educational researchers. By the very nature of the phenomenon we seek to understand we are forced to confront the problems associated with naturally occurring data arising from individuals nested within higher-order units. The problems inherent in such data are twofold. First is the difficulty of trying to explore causal issues through the use of non-experimental data. Second is the question of identifying appropriate statistical tools given some partial resolution to the first difficulty.

While the first of these problems is somewhat secondary to our purpose in this paper, a few words are called for. In the strictest sense causal inference cannot be drawn from non-experimental situations. We may, however, posit the existence of causal relationships among the variables in a particular situation and, from those relationships, deduce testable hypotheses in the form of a mathematical model. If our hypotheses consistently fail to be supported we are led to the conclusion that they or the process leading to them must be false. On the other hand, if our hypotheses are consistently supported, our causal presumptions are strengthened.

What then are the types of statistical tools applicable to such an effort? For the most part they are the very ones which we matter of factly apply to data arising from true experimental designs. We need but show appropriate caution in the interpretation of their results. Yet the mathematical models of greatest interest are based upon considerably more interrelationships than customary approaches have previously addressed. While we regularly deal with the relationship between one variable and another via correlational analysis, and between one variable and many via regression analysis, we have been poorly equipped to deal simultaneously with the interrelations of many variables. Path analysis has proven to be a fairly meaningful approach to this problem; however, its typical operationalization is structured around the independent estimation of relationships between one variable and many via regression analysis. Following such an approach is less than optimal for several reasons. We cannot constrain the values of the parameters in the model except in a very limited sense; measurement error is intolerable for all variables which appear as predictors of others; finally, no overall test of the adequacy of the model as a whole is offered by such an approach.

The technique which appears to come to grips with the level of complexity called for is one developed by Karl Jöreskog based upon his earlier work with confirmatory maximum likelihood factor analysis and the work in the analysis of covariance structures carried out at the University of Chicago by the second author of this paper, Dave Wiley, and others. The structural equation approach allows for the simultaneous estimation of interrelationships among latent or "true" variables through the use of maximum likelihood estimation assuming an underlying multivariate normal distribution. In addition, it is possible to place a number of constraints on the elements in the model. They may be free to vary, may be constrained to be equal to other parameters in the model, or they may be fixed equal to some arbitrary constant which may or may not be zero. Finally, the approach allows for both the estimation of standard errors for the parameter estimates as well as an overall test of the fit of the estimated model.

While this approach is a far cry from the examination of simple correlations between two variables, a very important question remains unanswered: What are we to do about the hierarchical nature of our data? One simple possibility is to ignore it and apply the structural equations approach to the overall variance-covariance matrix among the variables. This approach is fraught with dangers, not the least of which is the possibility that relationships that hold at one level do not hold at another. This means that we would be running the risk of estimating parameters that are themselves fundamentally meaningless if we were to adopt such a strategy. Alternatively, we may extract the within- and between-groups variance-covariance matrices and individually subject them to our structural equations approach. The drawback here lies in the fact that we are, once more, separately estimating parts of our model and lack a cohesive approach to the formulation of standard errors and tests of fit.

The solution lies in a formulation of the problem of estimating within and between groups parameters via maximum likelihood set forth by Schmidt (1969) in the context of the analysis of covariance structures. The approach taken there was one in which the observations were seen as being groupwise independent and multivariate normally distributed. Thus, the natural relatedness of subjects within groups was allowed to fully express itself in the

formulation of the likelihood function. The effective part of the logarithm of the likelihood function is expressed as:

$$\log L = \frac{m-mn}{2} \log \Sigma - \frac{m}{2} \log \Sigma + n\Sigma_a - \frac{mn}{2} \text{tr} \{\Sigma^{-1}S\} - \frac{m}{2} \text{tr} \{(\Sigma + n\Sigma_a)^{-1}S_a\}$$

Where

n = number of observations within each group

m = number of groups

Σ = within level covariance matrix estimated by the model

Σ_a = between level covariance matrix estimated by the model

S = within level observed covariance matrix

S_a = between level observed covariance matrix

The general form of the model specifying Σ and Σ_a closely parallels that developed by Jöreskog and is set out fully below. To facilitate the presentation, a notational convention has been adopted whereby a variable or parameter associated with the subjects within groups level stands alone and the corresponding parameter at the between groups level is subscripted with a lower case b.

The model may be set forth as two related components: the measurement model and the structural model.

Measurement Model

$$y = \mu + \Lambda \eta + \Lambda_b \eta_b + \varepsilon + \varepsilon_b$$

$$x = v + \Gamma \zeta + \Gamma_b \zeta_b + w + w_b$$

The μ and v vectors are simply the expected values of y and x respectively and are conceptually the same as the corresponding terms in Jöreskog's model. The matrix Λ contains coefficients relating the latent endogenous within-groups variables η to the observed variables, y . Likewise, Λ_b serves to relate the true endogenous between-groups variable, η_b , to the observed y . The vectors ε and ε_b represent the errors of measurement associated with within- and between-group levels respectively. The coefficient matrices, Γ and Γ_b , and the vectors ζ , ζ_b , w , and w_b bear similar relationships to the observed x vector.

Structural Model

$$\eta = A\eta + B\xi + \theta$$

$$\eta_b = A_b\eta_b + B_b\xi_b + \theta_b$$

The first equation stipulates that the latent within-groups endogenous variables are expressible as linear functions of themselves (as determined by the coefficients in the A matrix) and the latent within-groups exogenous variables (as determined by the coefficients in the B matrix). Finally, we have the vector θ containing the estimation errors associated with this part of the structural model. The second equation is composed of parallel constructs dealing with the expression of the between-groups latent endogenous variables.

As with Jöreskog's model there are a number of variance-covariance terms associated with these vector-valued variables. They are as follows:

- Σ_{ξ} - the variance-covariance matrix of the latent within-groups exogenous variables, ξ
- Σ_{θ} - the variance-covariance matrix of the within-groups errors of estimation, θ
- Ψ_{ε} - the variance-covariance matrix of the within-groups measurement error associated with the observed endogenous variables, y
- Ψ_w - the variance-covariance matrix of the within-groups measurement error associated with the observed exogenous variables, x
- Σ_{ξ_b} - the variance-covariance matrix of the latent between-groups exogenous variables, ξ_b
- Σ_{θ_b} - the variance-covariance matrix of the between-groups errors of estimation, θ_b
- Ψ_{ε_b} - the variance-covariance matrix of the between-groups measurement error associated with the observed endogenous variables
- Ψ_{w_b} - the variance-covariance matrix of the between-groups measurement error associated with the observed endogenous variables.

If we assume the measurement errors to be uncorrelated with each other and with the latent variables and that the errors of estimation are uncorrelated with the true exogenous variables and that all variables at one level are uncorrelated with those at another, the variance-covariance matrix of the observed variables can be expressed as a function of the parameters defined above.

Let Σ_z represent this matrix when $z' = (x', y')$. Then we have

$$\Sigma_z = \Sigma + \Sigma_b$$

and

$$\Sigma = \left[\begin{array}{c|c} \Gamma \Sigma_{\zeta} \Gamma' + \Psi_w & \text{symmetric} \\ \hline \Gamma \Sigma_{\zeta} B' (I-A)^{-t} \Lambda' & \Lambda [(I-A)^{-1} B \Sigma_{\zeta} B' (I-A)^{-t}] \Lambda' + \Lambda [(I-A)^{-1} \Sigma_{\theta} (I-A)^{-t}] \Lambda' + \Psi_{\varepsilon} \end{array} \right]$$

$$\Sigma_b = \left[\begin{array}{c|c} \Gamma_b \Sigma_{\zeta_b} \Gamma_b' + \Psi_{w_b} & \text{symmetric} \\ \hline \Gamma_b \Sigma_{\zeta_b} B_b' (I-A_b)^{-t} \Lambda_b' & \Lambda_b [(I-A_b)^{-1} B_b \Sigma_{\zeta_b} B_b' (I-A_b)^{-t}] \Lambda_b' + \Lambda_b [(I-A_b)^{-1} \Sigma_{\theta_b} (I-A_b)^{-t}] \Lambda_b' + \Psi_{\varepsilon_b} \end{array} \right]$$

where the partitions divide the matrices into parts associated with x and y respectively.

Given this formulation of the model and the effective part of the log likelihood function, we may produce estimates for the elements of the various parameters. We first take the partial derivatives of the log likelihood function with respect to each element. These partials are equated to zero and simultaneous solutions for each parameter are sought. Owing to the complexity

of the equations it is not feasible to derive general estimating equations. Rather, we must seek numerical solutions to these equations for each application of interest.

A computer program utilizing a combination of the steepest descent technique and the method of Davidson-Fletcher-Powell to yield the numerical solution for the model has been written by the first author. In addition, the program is designed to provide a chi square test of fit of the model based on the likelihood ratio. An additional program, still under development, will generate standard errors of the estimates based on expressions for the expected values of the matrix of second partial derivatives derived by Schmidt (1969).

The remainder of this paper is devoted to an example of a model utilizing data from the National Longitudinal Study of the High School Class of 1972. Sponsored by the National Center for Education Statistics, this study is a large-scale survey project whose primary purpose is the observation of the educational and vocational activities, plans, aspirations, and attitudes of young people after they leave high school. The Educational Testing Service began full-scale base year data collection in the spring of 1972. Data from over 18,000 students from a national probability sample of more than 1,000 high schools was collected. Beginning in the fall of the following year, Research Triangle Institute initiated the first of three follow-up surveys of these same subjects. To date, over 13,000 subjects have responded to all of the instruments administered.

The particular variables drawn from this data base include sex, ethnicity, father's educational level, mother's educational level, hours on English and foreign language coursework, and reading and vocabulary scores. To facilitate the analysis, sex and ethnicity were recoded to binary valued variables. For sex, zero represented female and one represented male; for ethnicity, zero represented blacks and one represented white. Within- and between-school variance-covariance matrices were calculated involving these variables. These matrices are set forth in the appendix to this paper.

The model treated sex, ethnicity, father's educational level, and mother's educational level as observed exogenous variables. Hours of English and foreign language, together with reading and vocabulary, constituted the observed endogenous variables. Father's and mother's educational levels were treated as two observed measures of a latent variable of socio-educational status.

The hours of English and foreign language were seen as observed measures of a verbal-skill coursework variable. Likewise, reading and vocabulary scores both measured a verbal aptitude trait.

The latent exogenous variables of sex, ethnicity, and socio-educational status were hypothesized to have a causal relationship with both verbal-skill coursework and verbal aptitude. Similar models were assumed to operate at both the within-and between-schools levels. The non-error related components of the model are diagrammatically presented in Figure 1, while the general parameterization of the components of the variance-covariance matrices is set forth in Figure 2.

These variance-covariance matrices, together with the within- and between-level sample sizes, served as the input to the estimation program discussed previously. The estimation routine was halted after completing approximately 500 iterations through the Davidson-Fletcher-Powell procedure. While the solution had not yet converged to the precision required by the program (the sum of all first derivatives being less than .0005), it is felt that these intermediate estimates provide a reasonably good idea as to the nature of the fully converged estimates.

The intermediate results presented in Table 1 suggest some interesting differences and similarities in the models at the different levels. To begin, the parameters associated with the scaling matrices, Γ and Γ_b , are quite similar. At the within-groups level, the intermediate estimate has a value of .762, while at the between-groups level, the corresponding value is .643. On the other hand, the estimates of the variances of the latent exogenous variables are quite different at within- and between- school levels. The within-school latent variance of socio-educational status (2,352) is more than twice that of the between-school level (.919). The situation for sex is considerably more extreme with the parameter estimate at the within-school level being .259 and at the between-school level .014. The same holds true for the estimates of the measurement error of the observed exogenous variables, ψ_w and ψ_{wb} . Thus, as might be expected, the latent exogenous variation is considerably larger at the within-groups level.

In considering the remaining parameter estimates, the difference across the levels is generally more pronounced. The estimates of the within- and between-group parameters relating the exogenous and endogenous latent varia-

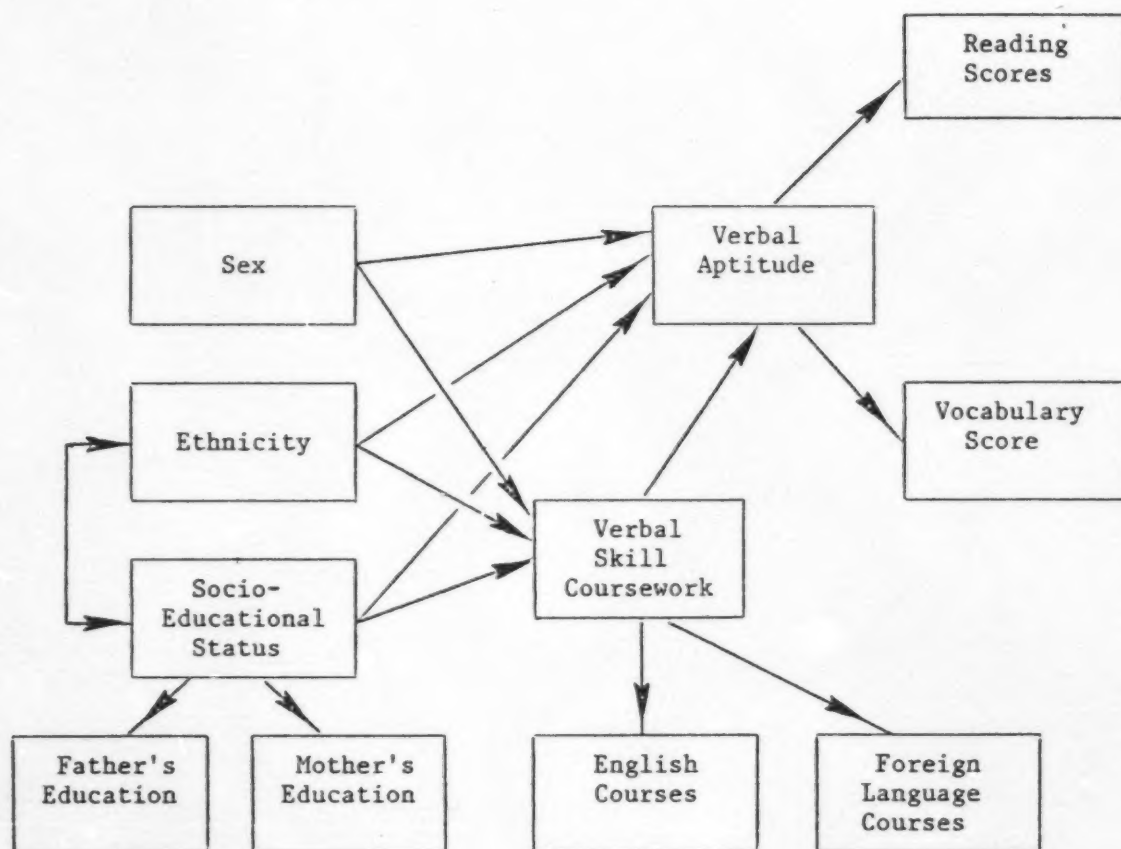


Figure 1. General Diagrammatic Structure of Estimated Model

Figure 2. General Parameterization of Model Variance-Covariance Matrices

$$\Sigma_{xx} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0_1 \end{bmatrix}}_{\Gamma} \underbrace{\begin{bmatrix} 0_2^* & 0 & 0 \\ 0 & 0_3^* & 0_4^* \\ 0 & 0_4^* & 0_5^* \end{bmatrix}}_{\Sigma_{\zeta}} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0_1 \end{bmatrix}}_{\Gamma'} + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0_6 & 0 \\ 0 & 0 & 0 & 0_7 \end{bmatrix}}_{\Psi_{\omega}}$$

$$\Sigma_{xy} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0_1 \end{bmatrix}}_{\Gamma} \underbrace{\begin{bmatrix} 0_2^* & 0 & 0 \\ 0 & 0_3^* & 0_4^* \\ 0 & 0_4^* & 0_5^* \end{bmatrix}}_{\Sigma_{\zeta}} \underbrace{\begin{bmatrix} 0_8 & 0_{11} \\ 0_9 & 0_{12} \\ 0_{10} & 0_{13} \end{bmatrix}}_{B'} \underbrace{\begin{bmatrix} 1 & 0 \\ 0_{14} & 1 \end{bmatrix}}_{[I-A]^{-1}}^{-t} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0_{16} \\ 0 & 1 & 0_{15} & 0 \end{bmatrix}}_{\Lambda'}$$

$$\Sigma_{yy} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0_{15} \\ 0_{16} & 0 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 0 \\ 0_{14} & 1 \end{bmatrix}}_{[I-A]^{-1}}^{-1} \underbrace{\begin{bmatrix} 0_8 & 0_9 & 0_{10} \\ 0_{11} & 0_{12} & 0_{13} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} 0_2^* & 0 & 0 \\ 0 & 0_3^* & 0_4^* \\ 0 & 0_4^* & 0_5^* \end{bmatrix}}_{\Sigma_{\zeta}} \underbrace{\begin{bmatrix} 0_8 & 0_{11} \\ 0_9 & 0_{12} \\ 0_{10} & 0_{13} \end{bmatrix}}_{B'} + \underbrace{\begin{bmatrix} 0_{17} & 0 \\ 0 & 0_{18} \end{bmatrix}}_{\Sigma_0} \underbrace{\begin{bmatrix} 1 & 0 \\ 0_{14} & 1 \end{bmatrix}}_{[I-A]^{-1}}^{-t} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0_{16} \\ 0 & 1 & 0_{15} & 0 \end{bmatrix}}_{\Lambda'} + \underbrace{\begin{bmatrix} 0_{19} & 0 & 0 & 0 \\ 0 & 0_{20} & 0 & 0 \\ 0 & 0 & 0_{21} & 0 \\ 0 & 0 & 0 & 0_{22} \end{bmatrix}}_{\Psi_{\epsilon}}$$

* These parameters were constrained equal to their unrestricted maximum likelihood estimates at the between-groups level of the model to enhance the properties of the estimation routine.

Table 1

Intermediate Parameter Estimates

<u>General Parameter Matrix</u>	<u>General Parameter</u>	<u>Within-school Level</u>	<u>Between-school Level</u>
Γ	θ_1	.762	.643
	θ_2	.259	.014
	θ_3	.108	.062
	θ_4	.084	.093
	θ_5	2.352	.919
Σ_{ζ}	θ_6	1.910	.008
	θ_7	1.590	.051
	θ_8	-5.042	-.775
	θ_9	2.323	-.378
	θ_{10}	4.680	.304
B	θ_{11}	.868	-.633
	θ_{12}	2.053	3.267
	θ_{13}	-.085	.908
A	θ_{14}	.209	.809
Λ	θ_{15}	.823	.903
	θ_{16}	6.107	179.138

Table 1 (continued)

<u>General Parameter Matrix</u>	<u>General Parameter</u>	<u>Within-school Level</u>	<u>Between-school Level</u>
Σ_{θ}	θ_{17}	166.412	.152
	θ_{18}	5.155	.379
	θ_{19}	6973.252	9284.864
Ψ_{ε}	θ_{20}	8.803	.343
	θ_{21}	5.429	.234
	θ_{22}	18297.432	.002

bles, B and B_b , are quite different in both magnitude and sign, indicating a difference across levels in these relationships. The latent endogenous variation unexplained by the latent exogenous variables, Σ_θ and Σ_{θ_b} , likewise differs markedly. Considering these differences, the relationship between the latent endogenous variables as indicated in A and A_b are quite similar in both direction and overall magnitude. The very pronounced difference between Λ and Λ_b , relating the latent endogenous variables to the observed variables, appears to indicate that the relationship between hours of English and hours of foreign language is somewhat different at these two levels. Given that such differences exist and given that English is usually required while foreign language is not, it would probably have made more sense to have reversed the roles of those two measured variables in defining the latent endogenous variable of verbal-skill coursework. Finally, the measurement errors estimated for the observed endogenous variables, Ψ_w and Ψ_{wb} , differ as well, as might be expected from the difference in the variability of the variables at the two levels.

When the remaining programming is finally completed, the facility will exist to both test the overall fit of the estimated model and generate asymptotic standard errors for the estimates. This work will soon be completed.

The preceding material has outlined and illustrated the authors' approach to the analysis of hierarchical data through the use of structural equation modeling. Further work is planned to evaluate this procedure in comparison to other methods for analyzing the same type of data. It is anticipated that the results obtained thereby will provide a strong argument for the adoption of this approach in the analysis of hierarchical data.

REFERENCES

- Duncan, O. D. Introduction to structural equation models. New York: Academic Press, 1975.
- Jöreskog, K. G. A general method for the analysis of covariance structures, Biometrika, 57(2), 1970, pp. 239-251.
- Jöreskog, K. G. A general method for estimating a linear structural equation system. In A. S. Goldberge and O. D. Duncan (eds.). Structural equation models in the social sciences. New York: Seminar Press, 1973, pps. 85-107.
- Jöreskog, K. G. and Van Thillo, M. LISREL: A general computer program for estimating a linear equation system involving multiple indicators of unmeasured variables. Princeton, N. J.: Educational Testing Service, December, 1972.
- Kerlinger, Fred N. and Pedhazur, Elazar. Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston, Inc., 1973.
- Schmidt, William H. Covariate structure analysis of the random effects model, Unpublished doctoral dissertation. University of Chicago, 1969.

APPENDIX

SYX WITHIN

		1	2	3	4
Hours of English	1	6647.950			
Reading Test	2	42.695	22.305		
Vocabulary Test	3	38.929	11.669	14.615	
Hours of Foreign Language	4	1273.014	261.267	215.373	24678.005

SXY WITHIN

		Hours of English	Reading Test	Vocabulary Test	Hours of Foreign Language
Sex	1	-.136	-.005	-.032	-7.345
Ethnicity	2	1.043	.338	.250	3.303
Father's Level of Education	3	11.183	2.042	1.798	61.936
Mother's Level of Education	4	9.318	1.563	1.403	47.324

SXX WITHIN

		1	2	3	4
Sex	1	.239			
Ethnicity	2	.003	.099		
Father's Level of Education	3	.015	.082	3.935	
Mother's Level of Education	4	.019	.054	1.653	2.624

SYX BETWEEN*

		1	2	3	4
Hours of English	1	127347.900			
Reading Test	2	-233.232	65.025		
Vocabulary Test	3	-294.020	45.653	48.806	
Hours of Foreign Language	4	1780.193	1172.148	1155.997	123295.380

SXY BETWEEN*

		Hours of English	Reading Test	Vocabulary Test	Hours of Foreign Language
Sex	1	8.470	-.041	.073	-29.849
Ethnicity	2	-5.683	4.211	3.474	.8340
Father's Level of Education	3	-98.283	18.654	17.521	628.049
Mother's Level of Education	4	-39.123	12.666	11.256	401.771

SXX BETWEEN*

		1	2	3	4
Sex	1	.440			
Ethnicity	2	.054	.909		
Father's Level of Education	3	.106	1.219	16.206	
Mother's Level of Education	4	.138	.804	9.414	8.412

* These matrices are n (or 13) times larger than the variance-covariance matrix among the school-level means.